ACT-R/E: An Embodied Cognitive Architecture for Human-Robot Interaction

J. Gregory Trafton, Laura M. Hiatt, Anthony M. Harrison, Franklin P. Tamborello, II, Sangeet S. Khemlani, Alan C. Schultz
Naval Research Laboratory

We present ACT-R/E (Adaptive Character of Thought-Rational / Embodied), a cognitive architecture for human-robot interaction. Our reason for using ACT-R/E is two-fold. First, ACT-R/E enables researchers to build good embodied models of people to understand how and why people think the way they do. Then, we leverage that knowledge of people by using it to predict what a person will do in different situations; e.g., that a person may forget something and may need to be reminded or that a person cannot see everything the robot sees. We also discuss methods of how to evaluate a cognitive architecture and show numerous empirically validated examples of ACT-R/E models.

Keywords: Human-robot interaction, cognitive modeling, cognitive architectures

1. Introduction

Robotic architectures that support human-robot teamwork have been developed for a broad range of interaction (Sellner, Heger, Hiatt, Simmons, & Singh, 2006; Kortenkamp, Burridge, Bonasso, Schreckenghost, & Hudson, 1999). For example, some architectures focus on optimizing user interfaces (Kawamura, Nilas, Muguruma, Adams, & Zhou, 2003); others support various depths of interaction, varying from teleoperation to shoulder-to-shoulder interaction (Fong, Kunz, Hiatt, & Bugajska, 2006). Such systems provide a broad base for human-robot interaction techniques, but for the most part they all share one key assumption: humans act as perfect autonomous agents, without the characteristics that make them – for lack of a better word – human.

Humans, even highly trained ones, are unpredictable, error-prone, and are susceptible to mental states like fatigue (Reason, 1990). In this light, we believe that it is important that a robot understands what human teammates are doing not only when they do something "right," but also when they do something wrong. Our goal in this work is to give robots a deeper understanding of human cognition and fallibilities in order to make the robots better teammates.

To do this, we rely on the computational cognitive architecture ACT-R (Adaptive Character of Thought-Rational) (Anderson, 2007). A cognitive architecture is a process-level theory about human cognition. For the purposes of HRI, a cognitive architecture can imbue a robot with a model of

Journal of Human-Robot Interaction, Vol. 2, No. 1, 2013, Pages 30-55. DOI 10.5898/JHRI.2.1.Trafton

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
ACT-R/E: An Embodied Cognitive Architecture for Human-Robot Interaction	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Research Laboratory ,Washington,DC,20375**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a REPORT	b ABSTRACT	c THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**25**	

the mental state of a human teammate, which can be exploited in any number of ways. Our choice of the cognitive architecture ACT-R is both natural and compelling. ACT-R has a rich history in the cognitive science community, whose researchers are concerned primarily with understanding how the mind works and how people think, perceive, and act. The techniques used by cognitive scientists have been applied to characterizing and improving ACT-R's fidelity to the human mind, strengthening our position that ACT-R can help robots better understand how people think and behave. Additionally, ACT-R is particularly well-suited to model human mentality because of the emphasis it places on understanding the limitations of human cognition.

In this paper, we describe our embodiment of the ACT-R architecture, Adaptive Character of Thought-Rational/Embodied (ACT-R/E). ACT-R/E adapts ACT-R to function in the embodied world by placing the additional constraint on cognition that cognition occurs within a physical body that must navigate and maneuver in space, as well as perceive the world and manipulate objects. Using ACT-R/E, we are able to build better, more comprehensive models of human cognition and leverage these models to improve our robots' ability to interact with humans. In what follows, we introduce the ACT-R and ACT-R/E architectures, and discuss how best to evaluate such architectures. We provide empirically validated examples supporting our claim that robots that understand people are, ultimately, better teammates and more natural computational agents.

We take "human-robot interaction" (HRI) to mean, generally, that humans and robots cohabit the same task environment (whether actual or virtual) and that actions performed by one entity may have consequences for another entity. We propose novel methods for robots interacting with humans in domains where some representation and processing of humans' states are critical.

2. ACT-R and ACT-R/E Architectures

A cognitive architecture is a set of computational modules that, working together, strive to produce human-level intelligence. The modules are typically designed to emulate different components of human cognition and can be tightly or loosely based on what is known about human cognition and how the brain functions. A cognitive architecture executes cognitive "programs," also called cognitive models. There exist a variety of cognitive architectures which emphasize particular areas of cognition or fidelity to the human mind or brain. The Soar architecture, for instance, is focused on achieving human-level intelligence and so uses a modest set of building blocks to achieve intelligence, including different types of memories (procedural, semantic, episodic) and different types of learning (reinforcement, chunking, semantic learning, episodic learning). It also has the ability to run on both virtual and embodied agents (Lehman, Laird, & Rosenbloom, 2006; Laird, 2012; Newell, 1990). Soar is concerned more with high-level functionality than with low-level cognitive fidelity, which makes it less suited to predicting people's errors and limitations. EPIC (Executive-Process/Interactive Control) emphasizes the effects that perceptual and motoric constraints place on cognition, which ACT-R in large part adopts into its own perceptual and motor modules (Kieras & Meyer, 1997). Polyscheme shows how different AI algorithms can be combined to achieve human-level intelligence by focusing on the benefits of multiple representational, planning, and reasoning systems (Cassimatis, 2002; Cassimatis, Trafton, Bugajska, & Schultz, 2004). Our work is based on the cognitive architecture ACT-R (Anderson, 2007; Anderson et al., 2004). ACT-R is a cognitive architecture that is meant to model human cognition at the process level and to address how humans' limited-capacity brains can handle the information processing requirements of their environment. We use ACT-R because of its dual focus on integrating different cognitive capabilities and human limitations (Anderson, 2007).

2.1 Canonical ACT-R

At a high level, ACT-R is a hybrid symbolic/subsymbolic production-based system. Given declarative knowledge (fact-based memories) and procedural knowledge (rule-based memories), as well as input from the world (visual, aural, etc.), ACT-R decides what productions to fire next; these productions can change either its internal state (e.g., by creating new knowledge) or its physical one (e.g., by pressing a key on a keyboard). It makes the decision of what production to fire next based on a) *symbolic* knowledge, such as who was where at what time; and b) *subsymbolic* knowledge, such as how relevant a fact is to the current situation, or how useful a production is expected to be when fired.

ACT-R is made up of up several major components. First, it has several limited-capacity buffers which, together, comprise its *context*. Each buffer is backed by one (or more) theoretically motivated modules (e.g., *declarative*, *visual*, *aural*, etc.); in addition, there is the procedural module, which does not have any associated buffers. Each module represents a specific cognitive faculty and has been shown to have anatomical correspondences in the brain (Anderson, Albert, & Fincham, 2005; Anderson, 2007). ACT-R's theoretical account for perception and action is limited to how cognition utilizes them. Certain processes, such as the transformation of raw sensory information into perceptual information, or the execution of action commands, are outside the architecture's scope. Therefore, what is actually perceived or done depends upon the sensor systems and effectors that feed into ACT-R. The architecture merely constrains how that information is used in cognition. Next, we describe the modules in more detail; then, we describe ACT-R's theory of how they interact. In Section 2.2, we will describe how ACT-R/E extends ACT-R to address a number of issues with using canonical ACT-R in HRI.

2.1.1 Declarative Module The declarative module manages the creation and storage of factual knowledge, or *chunks.* In addition, it manages memories' subsymbolic information by constantly updating their *activation* values. Activation values are a function of how frequently and recently that chunk has been accessed, as well as the extent to which the current context primes it (i.e., *spreading activation*). Activation also has a small random noise component that represents the fact that people's memory can be noisy. Mathematically, a chunk's activation value represents the log-odds that the chunk will be needed in the current context. Cognitively, a chunk's activation represents how long an ACT-R model will take to remember a chunk, if it can even be remembered at all. The theory that underlies activation values has been empirically evaluated in numerous situations and across various tasks, and has been shown to be an astonishingly good predictor of human declarative memory (Anderson, Bothell, Lebiere, & Matessa, 1998; Anderson, 1983; Schneider & Anderson, 2011).

Given a request to retrieve a declarative chunk, such as a prior solution to a similar problem, the declarative module attempts to find the best matching and most active chunk. The amount of time it takes to retrieve that chunk is directly related to its activation; after the chunk is retrieved, it is made available in the retrieval buffer. It may be the case that no chunk is retrieved if all matching chunks have very low activations or if there is no matching chunk at all.

2.1.2 Procedural Module The procedural module is mostly analogous to the declarative module, except that it creates and stores procedural knowledge, or productions. Subsymbolic information for production rules is represented by an *expected utility*. Expected utilities are learned over time according to a temporally discounted reinforcement learning function (Fu & Anderson, 2006) and, like activation, also have a small stochastic component. More importantly, however, expected utility is influenced by rewards and punishments defined by the modeler based on task and theory. Rewards and punishments can occur within an ACT-R model based on both internal events (e.g., successfully completing a goal) and extrinsic motivations (e.g., payment for performance).

Finally, the procedural module provides algorithms for matching the current contents of the buffers to production rules so that the strongest match (modulated by its expected utility) is selected to fire, and then handles implementing the results of the rule's actions. Neurological evidence, such as data from functional magnetic resonance imaging (fMRI) studies, suggests that the basal ganglia implement procedural learning (Ashby & Waldron, 2000; Saint-Cyr, Taylor, & Lang, 1988), and theorists have proposed that these learning processes can be mapped onto the operations of the procedural module in ACT-R (Anderson, Qin, Sohn, Stenger, & Carter, 2003).

2.1.3 Intentional and Imaginal Modules The intentional and imaginal modules provide support for task-oriented cognition. The goal buffer (intentional module) is intended to contain chunks that are used to identify the model's current goal and provide place-keeping within the task. The imaginal buffer, in contrast, is meant to provide support for intermediate problem state representations, such as a carryover digit in an addition problem. Another difference between these two modules is that the goal module specifies that new chunks are placed in the goal buffer immediately following a request to do so; the imaginal module, however, imposes a slight delay to reflect the effort made in transforming a mental representation (Anderson, 2007; Gunzelmann, Anderson, & Douglass, 2004).

2.1.4 Visual and Aural Modules The visual module in ACT-R enables the architecture to see elements in the model's world. Borrowed largely from EPIC (Kieras & Meyer, 1997), the module supports pre-attentive visual searching, visual object recognition, object tracking, and limited pop out effects (Byrne & Anderson, 1998). It also imposes temporal constraints on visual search and object recognition in a principled, cognitively plausible way. Similarly, the aural module is able to pre-attentively search its auditory environment and recognize sounds and utterances. Once attended to, visual and aural percepts are represented as chunks in declarative memory and are subject to the declarative module's management of their symbolic and subsymbolic information.

Because ACT-R was originally designed to explain and match data from cognitive psychology experiments that were presented on a computer screen, the design of the visual and aural modules considers a computer to be the entirety of a model's perceptual world; e.g., the visual module can only see objects displayed on a computer screen, and the aural module can only hear spoken words (such as instructions coming out of computer speakers). This *screen-as-world* paradigm limits the scope of embodied modeling possible within canonical ACT-R. For example, one assumption of the *screen-as-world* paradigm is canonical ACT-R's visual location representation of (x, y) screen coordinates; this makes accounting for real-world perceptual issues like occlusion or egocentricism nearly impossible.

2.1.5 Temporal Module The temporal module allows the architecture to keep track of time in a psychologically plausible manner. An internal timer is started to keep track of the interval between events. The temporal module acts as a noisy metronome, increasing the noise and decreasing accuracy the longer the interval lasts (Taatgen, Rijn, & Anderson, 2007).

2.1.6 Manual and Speech Modules The manual module is concerned with providing computer-related effectors to the model. Specifically, the architecture's body model is limited to 10 fingers and two hands, which are used to manipulate a keyboard, mouse, and joystick in response to commands placed in the manual buffer. Commands can only be issued serially, but the basic movements provided by the architecture can be strung together into complex sequences for typing and GUI manipulation. To maintain cognitive plausibility, the manual module uses Fitts' law (Fitts, 1954) to make predictions about the time course of movements.

Similarly, the speech module accepts commands via the speech buffer and executes them via an external text-to-speech processor. The speech module also makes coarse predictions of utterance duration. Both manual and speech commands, however, are never actually stored either symbolically

or subsymbolically. This prevents specific motor or speech information from being part of the architecture – one cannot remember a motor command if it was never encoded. The lack of a declarative trace also makes learning and refining motor movements a significant challenge. Additionally, as with the *screen-as-world* paradigm in the perceptual modules, this aspect of these two motor buffers severely limits the richness of embodied modeling possible in canonical ACT-R. Mounting evidence suggests that there is a tight (subsymbolic) link between perception and action (Craighero, Fadiga, Umiltà, & Rizzolatti, 1996; Tucker & Ellis, 1998), but such effects are difficult to account for using canonical ACT-R's manual, speech, and perceptual modules.

2.1.7 Module Interaction Although each of the above modules is theoretically motivated and validated on its own, it is the complex interaction of the modules over time that provides ACT-R's true predictive and explanatory power. Take the example of meeting someone at a party and then attempting to remember his or her name later in the evening. An ACT-R model performing this task would need to not only attend to the person's face, voice, and clothing, but also bind those individual chunks to the name of the individual. The name would then need to be rehearsed several times so it would not be forgotten. When the model saw the person a bit later, it would need to take available cues (perhaps only the face and the clothing) and attempt to retrieve from declarative memory the name associated with those cues. Priming from contextual cues like the face, clothing, and the party itself would provide the name chunk a boost in activation, and the earlier rehearsal would allow the memory to become active enough to be remembered. Integrating all this information at the time of meeting and at the time of remembering provides both descriptive information about how a human could err in this situation (e.g., people who look similar or who were met at similar times or at similar parties may be easily confused) and predictive information (e.g., remembering someone's name after a long break without further rehearsal or weak contextual cues makes it unlikely his or her name will be remembered).

Through this integration, ACT-R is able to model a wide range of human, including empirically grounded accounts of visual attention (Reifers, Schenck, & Ritter, 2005), time perception (Taatgen et al., 2007), working and long-term memory (Anderson, Reder, & Lebiere, 1996; Altmann, 2000), skill acquisition and selection (Taatgen & Lee, 2003; Fu & Anderson, 2006), multi-tasking and task switching (Salvucci & Taatgen, 2008), language learning (Taatgen & Dijkstra, 2003), interruptions and errors (Trafton, Altmann, & Ratwani, 2011; Trafton, Jacobs, & Harrison, 2012), and work load and fatigue (Gunzelmann, Byrne, Gluck, & Moore, 2009).

2.2 ACT-R/Embodied

ACT-R/E, shown in Figure 1, was developed to address a number of issues in canonical ACT-R that were a hindrance to embodiment, such as its *screen-as-world* world paradigm and its contextually limited perceptual and motor modules. To address these concerns, ACT-R/E makes a number of theoretical changes to canonical ACT-R. Chief among its theoretically motivated changes, ACT-R/E introduces two new modules to enable spatial reasoning in a three-dimensional (3D) world and makes modifications to its perceptual and motor modules to allow the tight linkage between perception and action.

Keep in mind as the discussion unfolds that ACT-R/E's strength is modeling cognition, not optimality. For example, the spatial module we describe below would not be appropriate for nor have the accuracy and resolution of state-of-the-art simultaneous localization and mapping (SLAM) algorithms (Montemerlo & Thrun, 2003). A SLAM algorithm, however, while useful to a robot during its own localization and navigational process, is not at all useful when trying to understand why humans systematically think they are closer to landmarks than they really are. This distinction is at the heart of our construction of embodied ACT-R.

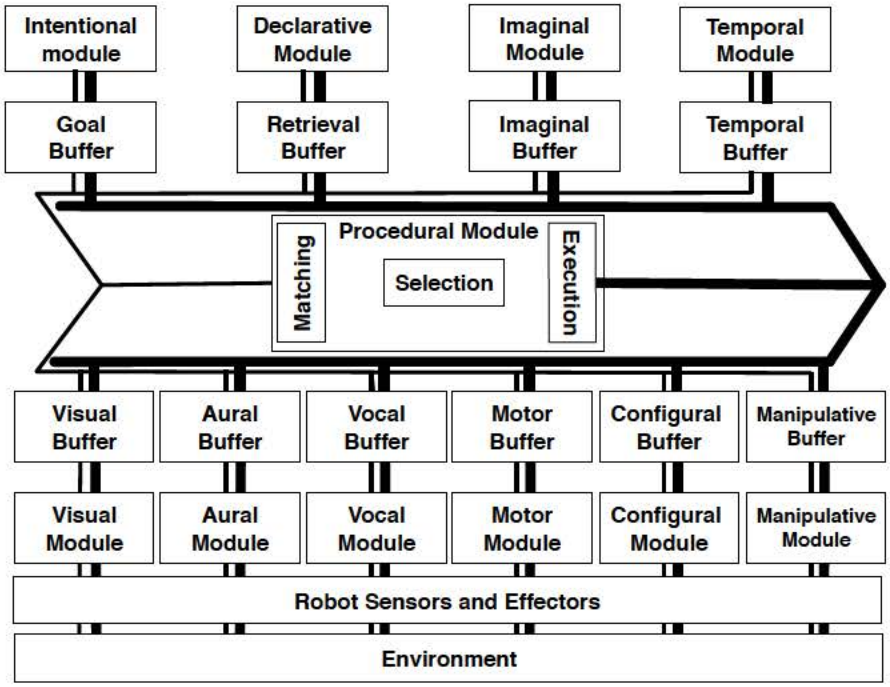

Figure 1. Architectural diagram for ACT-R/E.

2.2.1 Spatial Perception Spatial cognition is a core component of embodied cognitive models because it is a critical aspect of embodiment. Picking up objects, remembering where one's keys are, tracking people as they move around, and searching for a light switch all require fundamental spatial competency. To allow ACT-R to perceive and operate within real or simulated three-dimensional environments, the theoretical framework of Specialized Egocentrically Coordinated Spaces (SECS, pronounced "seeks") (Harrison & Schunn, 2002) was added into ACT-R/E, enabling human-like, cognitively plausible spatial reasoning.

SECS divides the spatial world into three regions, each of which is handled by a different module: the 2D-retinotopic space handled by the existing visual module; the configural space, for navigation and localization; and the manipulative space, to represent the graspable world. The partitioning of 3D space is neurologically inspired, with modular and functional boundaries coming from neuroimaging studies. However, the underlying process models of the functionalities were derived from behavioral data; see, for example (Mou, McNamara, Valiquette, & Rump, 2004; Jola & Mast, 2005).

The *configural* module integrates both visual and auditory information and represents their composition as egocentric vectors in space. This is used by a model to provide the information it needs to navigate above, around, under or over objects in the environment. ACT-R/E's motor module allows the configural system to update spatial representations during self-motion (Mou et al., 2004). In addition, the configural module allows for imagined transformations of the contents of the configural buffer (Mou et al., 2004; Jola & Mast, 2005), akin to what happens during complex perspective

taking. The imagined transformations are also critical for aligning ones perspective when reading maps. When attending to two configural representations (landmarks), the model is able to localize itself in space, relative to the two landmarks, giving the model the ability to recognize and return to unique locations in space.

The *manipulative* module, in turn, combines visual and motor information to represent objects in the immediately graspable space. Objects are comprised of an egocentric location vector, an orientation, and a metric, geon-based (Biederman, 1993) shape. These representations provide sufficient information for the model to be able to grasp and manipulate objects in the environment. Similar to the configural system, the manipulative module also allows imagined translations and rotations of the objects in the manipulative buffer (Jola & Mast, 2005).

2.2.2 Screen-as-World Paradigm As mentioned earlier, canonical ACT-R relies upon a *screen-as-world* paradigm, adapted from EPIC (Kieras & Meyer, 1997). This paradigm treats the computer screen and keyboard as the extent of the perceivable world. We have adapted both the motor and the visual modules in ACT-R/E so that models can perceive and move about in real-world situations.

When viewing the world, ACT-R/E's visual module is designed to act as a consumer of visual information provided by external visual systems. There are no restrictions on what visual systems can be used, nor on how many can provide information at a time. The same holds true for ACT-R/E's aural buffer and external auditory systems. Instead, the visual and aural modules constrain how that information is represented and used in cognition.

Recall that canonical ACT-R uses screen coordinates to represent visual locations. This is the result of an assumption of a static head and eyes, which a robot does not generally have, and limits the amount of reasoning about perceptual issues like occlusion or egocentricism that is possible. To address these limitations, ACT-R/E introduces a few modifications. First, all visual locations are retinotopic and not tied to any screen, allowing the robot to "see" more than a computer screen. Second, attention is object-based instead of location-based, allowing more sophisticated reasoning about occlusion and egocentricism. In addition, the visual system receives notifications anytime a potentially view changing movement (i.e., eye, head, torso) is going to occur so that it can compensate for the self-motion, allowing it to track moving objects while moving itself and to distinguish between object- and self- motion.

ACT-R/E also includes changes to canonical ACT-R's manual buffer. The manual buffer has a number of restrictions on it that are undesirable for robot use, including accepting only serial commands and timing movements with Fitts' law instead of tracking joint movements. To address these, ACT-R/E first extends the *manual* module into a richer *motor* module with a detailed and configurable internal body model, which allows us to keep the internal body model and actual physical capabilities in sync, and to better track movements as they are being executed. It also allows for parallel execution of non-conflicting muscles and limited proprioceptive feedback via muscle-specific queries.

2.2.3 Perceptual and Motor Integration One of the recent threads in cognitive science has been embodied, or grounded, cognition (Barsalou, 2010; Wilson, 2002). The focus has been on showing that the body has a major role in shaping the mind. Most of the work has included experiments that show that there is a tight link between the physical body and higher order cognition (Craighero et al., 1996; Tucker & Ellis, 1998). Until recently, however, there has been relatively little progress in formal or computational accounts of embodied cognition (Barsalou, 2010). One reason that it has been difficult to build computational accounts of embodied theories and phenomena is that they, by necessity, require a configurable body. In our opinion, robots are a promising way of filling this void (Pezzulo et al., 2011; Cangelosi & Riga, 2006; Tikhanoff, Cangelosi, & Metta, 2011).

Although there are other approaches that researchers are exploring – most notably the Perceptual

Symbol Systems Hypothesis (Barsalou, 1999) – our group is using an architectural approach to connect perception with action. The only way that perception and action can influence each other in canonical ACT-R is through symbolic representations engineered by the modeler (e.g., if the model sees a person, wave at the person). Our approach has been to allow perception and action to influence each other automatically through subsymbolic processes as well. Specifically, ACT-R/E allows activation to spread both to and from perception and motor chunks (Harrison & Trafton, 2010).

While ACT-R does enable any module buffer to be a source of spreading activation, the utility of that source depends critically on the module and its representations. ACT-R's manual module explicitly does not encode any declarative trace of its motor commands. Since the module does not represent the commands as chunks, not only can they not be encoded, retrieved or learned from, but they cannot spread activation to other associated chunks. ACT-R/E eliminates this constraint and allows motor commands to prime and be primed by other elements in the current context.

The visual module, in contrast, does create chunks that are ultimately encoded, and those chunks participate in the spreading of contextualized activation. However, visual representations are primarily composed of primitive features which, at best, severely reduce the contextual influence. Spreading of activation through the visual buffer is maximally effective if the model can associate the visual chunk with its symbolic meaning. This allows the perceptual chunk in the visual buffer to prime its symbolic partner. Instead of requiring the modeler to do this percept-symbol binding through productions, ACT-R/E handles this automatically upon attending the percept. The result is a visual percept chunk that automatically primes its symbol from the moment it is encoded.

With the motor and visual modules participating fully in the spreading of contextual activation, it is possible for a model to learn which objects are best grasped with which motor commands. As those associations are repeated, the two different representations can prime each other (Harrison & Trafton, 2010). Simply seeing a coffee cup will be sufficient to speed up access to the appropriate motor command needed (e.g., Tucker & Ellis, 1998, 2001). With both representations primed, all the chunks associated with coffee and the grasp will be boosted.

As we have described in the preceding section, ACT-R/E is able to build upon all of the canonical architecture's theoretical successes while simultaneously moving toward a richer account of embodied phenomena. Next, we discuss how we can apply these successes to HRI.

3. Examples and Evaluations of Cognitive Models

As capable as humans are in many situations, they are also susceptible to a variety of errors, mental states like fatigue, and unpredictability. In our view, there are three different ways a robot could handle such "human-like" behavior in teammates. First, the robot could assume that the human is always doing the "right" thing, even if it does not understand his or her behavior. This, however, can lead to more overall team errors (Ratwani & Trafton, 2009, 2011); such robots are also seen as less intelligent (Hiatt, Harrison, & Trafton, 2011). Second, the robot could continually question all of a human teammate's actions to make sure that the human is doing the right thing. Constant reminders, however, could lead to either the human ignoring them (Diaper & Waelend, 2000) or a less natural feel to the interaction (Hiatt, Harrison, & Trafton, 2011). Third, and the view we adopt as our own, the robot could be equipped with the functionality to understand the human's behavior, whether right or wrong, and use that information to act accordingly. We will show, in various scenarios, that this results both in humans making fewer overall errors (Ratwani & Trafton, 2011), as well as in more natural interactions, and in robots that are perceived as being more intelligent (Hiatt, Harrison, & Trafton, 2011).

Our high-level goal, then, is to give robots a deep understanding of how people think at the process level in order to make them better teammates. There are many possible implications of this

for HRI. First, people typically make errors in predictable ways. If a robot can understand how, for example, a person's eye gaze correlates to what they are thinking, they can use that knowledge to predict what they will do (or, sometimes more importantly, what they will not do) next. Second, people have imperfect memories. Knowledge of how memories decay over time, or how many things a person can remember at once, can help a robot to assist a teammate in situations where the person might forget something critical to the task. Additionally, if a robot incorporates process level models of different routes of thought that a person can take at any given time, they will better predict the person's behavior and ultimately be a better, and more helpful, teammate.

In some respects, our work on ACT-R/E has a dual nature. Our development of a cognitively-plausible embodied architecture, as well as of embodied cognitive models that work within that architecture, can be viewed both as furthering our work in HRI (by ensuring that the cognitive models that robots have are accurate models of people) and as pushing the boundaries in the field of cognitive science (by providing greater understanding of how people think and interact with the world).

In sum, our modeling approach can be characterized as a two-step process: (1) we develop embodied models on our robot that further our understanding of how people think, and (2) we leverage these models as tools for robots to use as they encounter humans in the world. In the following sections we first provide some thoughts on how to evaluate embodied cognitive models. Then, we provide short reviews of how ACT-R/E has been used in our laboratory, what the theoretical advances were, and how the models were evaluated. Most importantly, we describe how the validated cognitive models have been used as tools to enhance human-robot interaction.

3.1 Evaluating Cognitive Architectures

As described above, a cognitive architecture is a computational tool for a cognitive modeler to use to describe the cognitive processes a person goes through as he/she executes a task. As such, cognitive architectures are instances of Turing complete programming languages (Hopcroft & Ullman, 1979), i.e., they are computationally equivalent to universal Turing machines. Therefore, there is at present no way to evaluate an architecture as a whole, since it can be used to simulate any dataset or any functionality that is computable.

Cognitive models, however, can be evaluated, although there can be controversy over the means of model evaluation (Roberts & Pashler, 2000). Many cognitive models are evaluated using a match to empirical data – reaction time, accuracy, or the time course of neurological data (e.g., galvanic skin response or brain regions or event-related potentials). Cognitive models can also be evaluated by showing novel ability (reasoning as well as a person does), breadth (applying to different situations), and simplicity (providing a parsimonious set of mechanisms) (Cassimatis, Bello, & Langley, 2008).

To evaluate our architecture, then, we test it in three different ways: (1) we test and evaluate each component separately, to validate it against human subject data; (2) we test different sets of the components as they interact; and (3) we show how our models increase the ability, breadth, and parsimony of cognitive models. Together, we consider these tests to serve as proxy metrics for the architecture as a whole. As we described each component of ACT-R, we mentioned some of the tests and supporting evidence that provide them with theoretical justification. In this section, we focus on the second type of evaluation. Below we describe four models that test different sets of ACT-R/E's key components. Table 1 summarizes the applications, the components that were critical to the model's success, and the dataset against which the cognitive model was assessed.

In all the examples below, a computational cognitive model was built and matched to human data. Model runs usually occurred in a simulated environment. After (and frequently during) the building of the model, the model was then run on an embodied robot. Many advantages come from running the cognitive model on an embodied platform. First, the running model showed that the

Task	Components of ACT-R/E	Dataset
Gaze following	Manipulative module	Corkum & Moore (1998)
	Configural module	Moll & Tomasello (2006)
	Utility learning	
Hide and seek	Imaginal module	Trafton, Schultz, Perzanowski, et al. (2006)
	Visual module	
	Vocal module	
Interruption and resumption	Declarative module	Trafton et al. (2012)
	Intentional module	
	Imaginal module	
	Procedural module	
Theory of mind	Declarative module	Leslie, German, & Polizzi (2005)
	Architecture as a whole	Wellman, Cross, & Watson (2001)

Table 1: Human-robot interaction tasks to which the ACT-R/E cognitive architecture has been applied

interaction was appropriately represented in its embodied form and that the overall system behaved as expected (this was done for all of the gaze-following, hide and seek, interruptions, and theory of mind models). Second, because most of the models were based on (typically single or simplified) situations, the embodied robot showed that the model was general enough to go beyond that single situation or task (gaze following, hide and seek, interruptions, theory of mind). Third, because most of the models were built based on a specific set of human participants, running the model in an interactive scenario showed that the model was not limited to the set of participants the model was built for (hide and seek, interruptions). Fourth, it is critical to perform some sort of evaluation to show that the interaction itself was successful and that people could understand and deal with the robot (hide and seek, interruptions, theory of mind). Fifth, putting the model on an embodied platform highlights the integration between components, showing that intelligent behavior requires multiple capabilities, all acting together (gaze following, hide and seek, interruptions, theory of mind). Finally, running the models on an embodied platform forces the model to go beyond the simplified view of reality that simulations provide (Brooks & Mataric, 1993) (gaze following, hide and seek, interruptions, theory of mind).

3.2 Gaze Following

First, we explored how very young children (6–12 month olds) learn to follow the gaze of others. Previous research has shown that reliable gaze following typically begins around 11 months of age and is an important cognitive competency for engaging with other people (Corkum & Moore, 1998). We built an ACT-R/E model that simulated the development of gaze following and showed not only that gaze following is a learned behavior (not a solely inborn one), but also how ACT-R/E's spatial modules interact with each other to bring about gaze following.

3.2.1 Model Description The ACT-R/E model had three components that were critical to the success of the model: the manipulative module, gaze extrapolation, and utility learning. As discussed above, the manipulative module provides information about what direction an object (in this case, a head) is facing. Gaze extrapolation then takes the heading of the face via the manipulative module and performs a directed visual search from the eyes in the direction that the head is pointing; objects can be found along this line. ACT-R/E's utility learning mechanism is used any time a reward is

given (e.g., for infants, a smile from a caregiver); i.e., the reward is propagated back in time through the rules that had an impact on the model getting that reward.

When the model was "young," it had a favored rule set, which was to locate, attend to, and gaze at an object. The model would look at objects but, because infants enjoy looking at faces, the model would return to look at the caregiver after looking at an object. The object could be anything in the model's field of view and was chosen randomly. If the caregiver was looking at the same object that the model decided to look at, the model was given a small reward. If the caregiver was looking at a different object than the model, no reward was given but the trial completed and the reward process began anew. In both cases, because the model returned to look at the caregiver after looking at the object, it could see the caregiver's response.

This young model, however, also had an unfavored gaze-following rule available to it. The gaze-following rule had a much lower utility when the model was young, so it did not get an opportunity to fire very often. However, because of the relatively high noise value for utility, the gaze-following rule did occasionally get selected by the procedural module and fired. When the gaze-following rule fired, the infant model looked at the caregiver, determined where he/she was looking using the three step process described above, and then looked from the caregiver to the direction he/she was facing. The model then found the first available object in that direction, which is a process consistent with previous research (Butterworth & Jarrett, 1991). The model was given a small reward if it saw the caregiver smile. After the model got tired of looking at that object (habituation), the trial ended and the model looked for another object to attend to. Because the gaze-following production was rewarded more often than the random production, which was rewarded an average of $number\text{-}of\text{-}objects^{-1}$, the gaze-following production slowly gained utility. However, it took a period of time before the combination of noise and utility allowed the gaze-following production to overtake and eventually become dominant over the random-object production. Further details are available elsewhere (Trafton & Harrison, 2011).

3.2.2 Model Evaluation This model was evaluated by comparing the results of the model to previously collected human data (Corkum & Moore, 1998). Three age groups (6- to 7-, 8- to 9-, and 10- to 11-month-olds) completed the experiment. In the experiment, the objects were two toys, and trials started when the infant looked at the experimenter. The experiment consisted of three consecutive phases. In the baseline phase, the experimenter simply looked at a toy. During the baseline phase, the toy remained inactive (i.e., did not light up or turn) in order to assess spontaneous gaze following. In the shaping phase, regardless of the infant's gaze, the toy that was gazed at by the experimenter lit up and rotated. Then, during the final testing phase, the toy was activated only if the infant and the experimenter both looked at the same toy. Each head turn was coded as either a target (joint gaze with the experimenter) or a non-target (the wrong toy was gazed at) response. Random gaze following would correspond to approximately 50% accuracy, whereas accurate gaze following would be greater than 50%.

As Figure 2 suggests, only 10–11 month old infants could reliably follow gaze at baseline. After training, however, both 8–9 month and 10–11 month old infants could reliably follow gaze (there was a slight, non-significant increase in gaze-following for the 6–7 month old infants). The circles in Figure 2 indicate the model fit from ACT-R/E. In all cases the model is in the 95% confidence intervals of the data, showing an excellent fit ($R^2 = .95; \mathrm{RMSD} = 0.3$).

A similar model was also applied to a slightly different task: level 1 visual perspective taking, or understanding that what one sees can be different from what another person sees. The model for level 1 visual perspective taking used the same components as the gaze-following model. After learning, the model was able to perform a simple visual perspective-taking task. The model was evaluated by matching the ACT-R/E model data to previously collected human data (Moll & Tomasello, 2006);

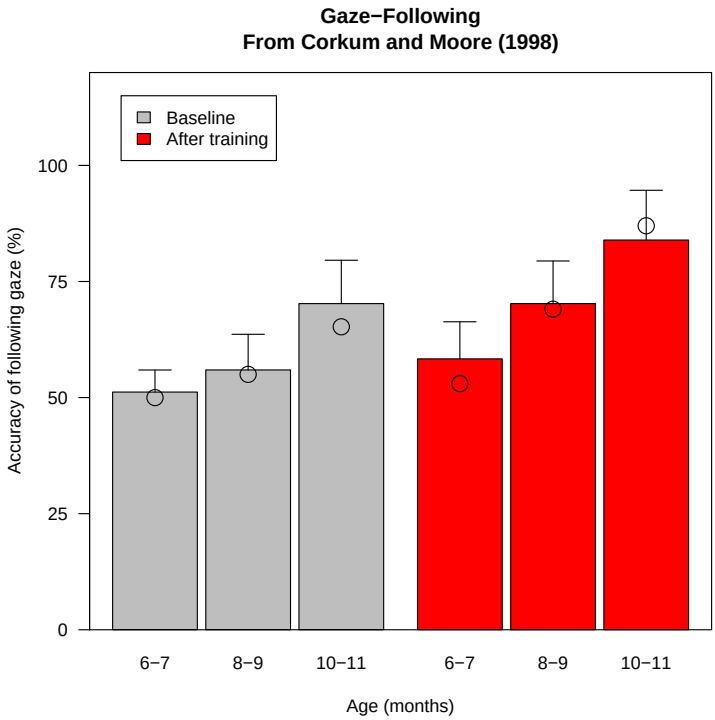

Figure 2. Experimental data from Corkum and Moore (1998). Bars are experimental data and circles are model data. Error bars are 95% confidence intervals.

$R^2 = .99; \mathrm{RMSD} = 10.2.$

3.2.3 Contribution to HRI The ability to follow gaze and show even very simple visual or spatial perspective taking has been shown to be an important aspect of HRI (Mumm & Mutlu, 2011; Satake et al., 2009; Szafir & Mutlu, 2012). Our MDS (Mobile, Dexterous, Social) robot (Breazeal et al., 2008) acquired this important skill by using the model to learn how to follow gaze. After learning, the robot was able to follow a person's gaze to objects in the room that it could see (Trafton & Harrison, 2011). A robot that is able to follow gaze shows that it is paying attention to people around it and shows some awareness about its environment. The ACT-R/E model of gaze following can also be compared to previous models of gaze following (Doniec, Sun, & Scassellati, 2006; Nagai, Hosoda, Morita, & Asada, 2003; Triesch, Teuscher, Deak, & Carlson, 2006). Previous models either did not have psychologically plausible spatial or attentional mechanisms or they were not embodied on a robot.

Gaze following has been considered to have strong social roots, but our approach shows that at least some social behaviors can be modeled using an embodied cognitive science approach. Additionally, using our embodied approach, the robot knows that people cannot always see an object that it itself can see: people can not see objects behind them, for example, nor can they see objects that are occluded. If the robot understands that people can not always see everything in the robot's view, the robot can use that knowledge to improve HRI by, for example, telling a user where a

needed wrench is if the robot can see it but the person can not (Trafton et al., 2005). A video of the robot performing gaze following is available at http://www.nrl.navy.mil/aic/iss/aas/movies/Gaze-following.10-11m.mov.

3.3 Hide and Seek

The second example of an embodied cognitive model is focused around the children's game "hide and seek." Hide and seek is a simple game in which one child, "It," stays in one place and counts to ten with their eyes closed while the other players hide. Once ten has been reached, the "It" child goes to seek, or find, the other child or children who have hidden.

Although the rules of the game itself are relatively simple, not all the cognitive processes needed to play hide and seek are simple. Spatial perspective taking is one cognitive ability required for a good game of hide and seek: a good hider needs to take into account where "It" will come into a room, where "It" will search first, and where to hide behind an object taking the perspective of "It" (Lee & Gamard, 2003) so that "It" will not be able to find the hider easily. Additionally, the hider must know that just because the hider cannot see "It" doesn't mean that "It" cannot see the hider.

Hide and seek, however, has been shown to have an apparent contradiction in this regard: young children (3 1/2 year-olds) do not have full spatial perspective-taking ability[1] (Huttenlocher & Kubicek, 1979; Newcombe & Huttenlocher, 1992), but they are able to play a credible game of hide and seek. By credible, we mean that 3 1/2 year-old children are able to find a place to hide that is not totally out in the open and that can be challenging for another 3 1/2 year-old (or adult!) to find them. We built an ACT-R/E model to explain this anomaly and provide a working hypothesis about how young children without perspective-taking abilities are able to learn to play hide and seek. The ACT-R/E model suggests that although young children do not have spatial perspective-taking abilities, they are able to learn simple relational rules which allow them to play a reasonable game of hide and seek. For example, a child may learn that hiding under or inside of an object would be a good hiding place. In contrast, hiding behind an object would be more difficult to successfully execute because the concept of "behind" requires some spatial perspective taking.

This hypothesis was supported by examining how a 3 1/2 year-old child, E, learned how to play hide and seek. E received feedback after every game ranging from "That was a good hiding place!" (positive feedback) to "I can still see you!" (slightly negative feedback). The child played 15 games of hide and seek and went from hiding out in the open with her eyes covered to hiding under a covered chair. In the majority of games, the child hid either inside an object (e.g., a box or a room) or under an object (e.g., a piano). In only one instance (7%) did E hide behind an object (a door), and that case could also have been construed as being inside a room. Further details are available elsewhere (Trafton, Schultz, Perzanowski, et al., 2006).

3.3.1 Model Description The ACT-R/E model provided an embodied cognitive process description of exactly how a child could learn to play hide and seek without having full spatial perspective taking. The model suggests that multiple learning mechanisms need to be integrated to learn how to play hide and seek. The model starts off with knowledge about the rules of the game, the ability to search for candidate "hide-able" objects, and some knowledge about spatial relationships (under, on top of, inside, etc.).

As the model played a game, it looked for a place to hide. Since it had no spatial perspective taking, it could not look for large opaque objects to hide behind. Instead, the model began each game by inspecting its world and looking for interesting objects. After it chose an interesting object, it attempted to hide near that object – inside of the object, under the object, on top of the object,

[1] As noted in the Gaze Following section above, young children are able to perform simple gaze following (by age 1 year) and level 1 visual perspective taking (by age 2 years).

etc. The object itself also provided affordances which informed the hiding location (e.g., it is easier to hide under a piano than on top of it). The model then received some feedback (similar to what E received) on whether that was a successful hiding place or not. That feedback was then used to reason about future hiding places and the model could learn what types of places were good or poor for hiding. Critically, the model learned that some features of an object were good for hiding (e.g., an object that had an inside was better than an object that only had an "on top of" relation) and some features were not (e.g., an object that could be seen through is not a good hiding place). The model thus used some simple reasoning strategies to learn about hiding places. After playing several games, the model learns a relatively stable set of "good" hiding places (objects) as well as good hiding features (e.g., inside).

3.3.2 Model Evaluation The hide-and-seek model was evaluated first by examining how closely the hiding model matched the hiding behavior of the 3 1/2 year-old child. Interestingly, the model was highly dependent not only on explicit feedback of a good or poor hiding place, but also on what objects were available to hide under or inside of. The model did a surprisingly good job of mimicking the outward behavior of the child, perfectly matching the hiding behavior (Trafton, Schultz, Perzanowski, et al., 2006).

The model was also put onto a robot (Coyote, a Nomad 200) so it could learn to play hide and seek in an embodied environment. The robot model was able to navigate around its environment and had some limited computer vision to detect objects in its environment (Trafton, Schultz, Perzanowski, et al., 2006). The robot was also able to accept commands ("Go hide") via natural language. When the robot started playing, it had some knowledge about objects (e.g., the size) and was able to reason about good or poor places to play, based on natural language feedback that the seeker provided. As the robot played, it was able to learn good and poor hiding places. Although no formal user evaluation was performed, the robot successfully learned how to play hide and seek with different visitors in the laboratory. Success in this case was operationalized as whether the visitor needed to look for the robot in more than one place.

The model was also evaluated by changing the environment itself. Because the model learned about features of objects, it was possible to change the environment by adding objects it had no experience with to see how well the model adapted its hide-and-seek game. In this situation, the system did not determine all features through perception but the model was provided with a few features of the objects — size and opacity (although not location). When put in a novel environment, the robot model was again able to play a credible game of hide and seek against a human.

The hide-and-seek robot model was also evaluated using another classic evaluation technique: generalization of the model to a new task. In this case, the model was generalized to play as "It," a different situation where similar, but not exactly the same, knowledge would be needed. The seeking model was able to use the hiding model's knowledge to look for a hider according to its own model of hiding. In other words, it searched in places that it determined were plausible for it to hide in. As expected, the embodied model systematically searched different locations that it had learned were acceptable hiding places until it found the person hiding. Again, no formal evaluation was performed, but the robot model was able to both hide and seek with different visitors to the laboratory. Success in this case was measured by the amount of time that (adult) visitors wanted to play with the robot: in some cases other demos had to be cut short because of the success of this demo.

3.3.3 Contribution to HRI The complete system – an embodied model that learned how to hide and then could use that knowledge to seek – is a novel aspect of HRI. The model not only describes the types of knowledge needed when a 3 1/2 year-old learns to hide, but it also was able to use that information to seek. It did not hide in a place that would be quite difficult to find for a young child

and it did not bother looking for a person in a place that a child would have a low likelihood of searching. One of the interesting aspects of this work is not only showing what type of knowledge and representations children use when playing hide and seek, but also showing that a model that learned how to hide could use that knowledge to seek.

The hide-and-seek model used the intentional module (to give it some top-down guidance), the declarative module (to store the results of games and reasoning processes), the imaginal module (to store intermediate steps in the hiding or seeking game), and the procedural module. The visual, aural, vocal, and motor modules were also used.

Note that for this example, the robot never actually interacted with a child. However, somewhat surprisingly, a robot that could play hide and seek at a 4 year-old's level could play hide and seek with adults. Informally, all the adults who played hide and seek with the robot were extremely engaged with the robot and the game itself, attempting to "trick" the robot. We had approximately 60 visitors who played or watched hide and seek with the robot. We also had several offers to take the robot home to play hide and seek with visitors' children. More formal examples and evaluations of our overall approach appear below (e.g., the interruptions and theory of mind examples), but it should be noted that this robot model was considered highly interactive to visitors who came into our laboratory. A video of the robot playing hide and seek is available at http://www.nrl.navy.mil/aic/iss/aas/movies/hideNseek-Feb06.mov.

3.4 Resuming After an Interruption

In today's connected world, interruptions and multi-tasking are a huge part of daily life. There have been several studies that have shown that interruptions are frequent (every 11 minutes; González & Mark, 2004), are time-consuming (consuming up to 28% of an office worker's time; Spira & Feintuch, 2005), are expensive (US $588 billion a year; Spira & Feintuch, 2005), and greatly increase the probability of errors (10x or more in some situations; Ratwani & Trafon, 2011; Trafton, Ratwani, & Altmann, 2011). Given the prevalence and impact of interruptions, building systems that can facilitate resumption after an interruption would have multiple benefits.

We built an ACT-R/E model that emulates the process people go through as they get interrupted and then have to resume their task. The ACT-R/E model matches human data, and thena robot model uses the ACT-R/E model to predict whether or not a person forgot what he/she was going to do; if the model predicts that the person did forget, it can then remind the person of the last thing he/she had done or said (Trafton et al., 2012). For this example, we explored how people resume after being interrupted while telling a story. The model was then matched against human-participant data.

The theoretical approach we take is based on the memory for goals theory (Altmann & Trafton, 2002, 2007; Trafton, Altmann, Brock, & Mintz, 2003; Trafton et al., 2011). Memory for goals provides a process description about how people remember, retrieve, and forget goals they have worked on. The theory posits that when a person begins work on a goal, he/she creates an episodic code that can be strengthened by rehearsal or associations with cues in either the physical (by looking at the world) or mental (by thinking about the goal) environment. This episodic code (also known as a control code) is a unique identifier that marks what the system is currently working on. The episodic code itself is part of declarative memory, although it has a relatively lean representation: elaborations to the episodic code occur in other parts of declarative memory and are typically much more robust. As another goal is started or the person gets interrupted, earlier codes lose strength and can become more difficult to remember. This forgetting process is functional – it keeps the system (human or computational) from getting clogged down with minutia. In the case of the interruptions model below, episodic codes can serve not only as a marker of what the system is currently working on, but also as a placekeeper when someone needs to remember what he/she was doing after being interrupted.

3.4.1 Model Description At each storytelling gist, an episodic control code is created. Which episodic codes are in declarative memory at any given time therefore marks the history of the storytelling. If a story is interrupted and needs to be resumed, then, at resumption time the model attempts to retrieve the most active episodic code, which is typically the most recent. If successful, it will use that code and the associated gist element to retrieve the next to-be-reported gist element, allowing the model to continue the task. If the model fails to retrieve a relevant episodic control code, one of two options is available to the model. If there is a listener available, the model will merely ask for help. If, however, there is no listener available or the listener is unable to help, the model will try again to remember an episodic tag and will make repeated attempts until it successfully retrieves an episodic tag. As a last resort, the story can be started over. Because there is noise associated with memories, the model sometimes does make mistakes. Interestingly, this model provides an explanation for how transactive memory occurs: the listener can serve as another memory source, but is only used if the storyteller lost his/her place. Note that the model itself not only is a theory of resuming after an interruption, but also can be used to predict when someone has forgotten where he/she was after being interrupted.

3.4.2 Model Evaluation We evaluated our model by comparing its performance to empirical data we collected (Trafton et al., 2012). In that experiment, 39 participants memorized a soap opera-like story and then retold the story to a confederate or to a video camera. Half of the storytellers were interrupted while retelling the story and then had to resume after the interruption finished; the other half served as a control condition to examine uninterrupted performance. After the interruption, storytellers attempted to retell the story where they left off; the confederate was not helpful at all. We recorded whether each person asked for help, how long it took for the person to resume, and where they resumed. The ACT-R/E model was quite accurate at emulating human performance, coming within 3% of the number of times that people needed help (corresponding to the model being unable to retrieve any episodic code), being very close to the amount of time it took people to resume (RMSD = 1.5), and matching the pattern that people were more likely to resume earlier on in the story than skip steps, consistent with the memory for goals prediction (Trafton et al., 2011).

3.4.3 Contribution to HRI We next used our model of interruption recovery to improve human-robot interaction by running the model on our MDS robot in the place of the confederate. Twenty-two participants were run in the robot study: 11 in the responsive condition with the model and 11 in a control condition that did not attempt to help. One critique that is sometimes made of highly controlled empirical studies is that the participants are frequently college students, female, and under 25 years of age, which is not representative of the rest of the population. For this robot experiment, we attempted to generalize the model to a different group of participants, which is quite important for model validation. All the participants in the human-human experiment were women, while their average age was 20 years; half of the participants in the human-robot experiment were women, while their average age was 42 years. The procedure with the robot was identical to the human-human experiment. At the time of interruption, if the model predicted that the storyteller had forgotten where she was, the responsive robot told the storyteller the last thing she/he had said before the interruption. The model was highly successful: it helped 100% of the people who seemed to need help and had relatively few false alarms. We also examined how the model would have performed if the model had been run on the control condition. We found that over 70% of the participants who needed help would have received help from the model and that no one would have received help if they did not need it. Interestingly, we found that participants who received help thought the robot was more natural and more useful (both p's < 0.05) than when the robot did not provide help.

This model leans heavily on the declarative module of ACT-R/E, although the intentional mod-

ule (for goals) and imaginal module (for problem representation) were also used. The procedural module was used as well.

This model made two contributions to HRI. The first contribution is that the model allows a robot to accurately predict whether a human teammate will remember something after an interruption. The second contribution is that we have shown how a robot can leverage this knowledge to improve human performance while also being perceived as natural and useful. This model and the associated study also highlight our overall point of the fallibility of humans, and how modeling human limitations can be a powerful tool in human-robot interaction.

3.5 Theory of Mind

Theory of mind, or the ability to understand the beliefs, desires and intentions of others, is a critical capability for teams of agents (whether human or robot) working together to accomplish a task. It is a complex, heavily-studied cognitive phenomenon not only in cognitive science (Hiatt & Trafton, 2010; Friedlander & Franklin, 2008) but also in developmental psychology (Wellman et al., 2001) and cognitive neuroscience (Gallese & Goldman, 1998). From this work, several competing theories have been formed about how theory of mind, or ToM, develops as children grow. One of the main divisions in the community is whether development occurs as a result of learning ToM concepts and causal laws (for convenience, we will refer to this as "learning"), or, alternately, as a result of increasing capabilities and functionality of cognitive mechanisms in the brain (we refer to this as "maturation"). Two cognitive mechanisms typically discussed are the ability to select between candidate beliefs, and the ability to perform cognitive simulation.

These two competing views have been extensively discussed in conjunction with two developmental shifts in ToM: one, which occurs at about 3–4.5 years of age, when children go from being mostly incorrect to mostly correct on a simple ToM task which involves identifying beliefs of others; and a second, which occurs at around 4.5–6 years of age, when children gain the ability to correctly perform ToM tasks that involve using others' identified beliefs to predict others' behavior (Leslie et al., 2005). We have developed a cognitive model for ToM that posits a cohesive theory of how children develop the ability to perform both of these ToM tasks (Hiatt & Trafton, 2010). Then, we loaded it on the robot to improve HRI (Hiatt, Harrison, & Trafton, 2011).

3.5.1 Model Description Our cognitive model of theory of mind proposes that children learn and mature simultaneously. The model supports this theory by developing ToM as it "ages" in the world. There are three mechanisms that allow this to be accomplished: (1) ACT-R's utility learning mechanism; (2) a belief selection mechanism; and (3) a cognitive simulation mechanism (Kennedy, Bugajska, Harrison, & Trafton, 2009). ACT-R's learning mechanisms have already been described. The belief selection mechanism allows a model to select between competing beliefs to select something other than the most salient belief, similar to Leslie's selection by inhibition (Leslie et al., 2005). The cognitive simulation mechanism allows the model to spawn a simulation by using its own decision-making systems to operate on the identified mental states of others to predict the others' behavior. During the cognitive simulation, the model spawns a submodel with: the other's identified belief, access to the model's productions and modules, and the (explicitly stated) goal of the other, and returns what it thinks the other person will do next.

Our ToM model starts out at approximately 2 years of age with the ability to generate different possible beliefs that people may have, but no mechanism to select between them. During the task, it sees two children in a room with a marble and two boxes. Next, it observes child "A" put a marble in box 1 and then leave the room. Then, it sees child "B" move the marble to box 2. Finally, it sees child A return to the room and is asked where child A thinks the marble is. The model is rewarded if it gets the answer correct. Since it has no ability to distinguish between the different possible beliefs

of the marble's location, it will always fire the production that leads to thinking that others have the same beliefs as it does and will fail even the simple ToM task. As the model repeats this ToM task over time, however, it slowly begins to repress its knowledge of the actual location of the model in order to select the location of the marble as child A saw it. The production that performs this belief selection is fired intermittently due to utility noise; therefore, once the model starts to gain this ability, it also simultaneously uses ACT-R's utility learning mechanism to learn *when* it should select between them, leading to eventual success on the simple ToM task.

Then, at around 4–5 years of age, the model slowly gains the ability to perform the more difficult ToM task. During this task, the marble is replaced by a kitten that crawls between the boxes while child "A" is out of the room. When the child returns, she wants to put a piece of fish under the unoccupied box, and the model is rewarded if it can correctly identify where child A will try to put the fish. To predict child A's behavior, the model first identifies her belief as in the simple ToM task. Then, it uses cognitive simulation to use that belief to predict where the child will put the fish. At first, simulations are likely to fail, but as the model ages, simulations become likely to succeed and the model learns that it is a good way of solving the problem. As with the belief selection mechanism, the model then simultaneously develops the functionality to perform simulation, and learns how and when to do it.

Figure 3. Screen shot from a scenario where the MDS robot and a human teammate are patrolling an area.

3.5.2 Model Evaluation We evaluated our model by running various simulations of the model learning how to perform the two ToM tasks (Hiatt & Trafton, 2010). The model was asked to perform the tasks repeatedly as it "aged" and was either rewarded or punished based on whether it performed the task correctly. Simulations were utilized in order to enable fast model development and data collection; to maintain fidelity with our embodied goals, simulations were run within the open-source, robotic simulation environment Stage (Collett, MacDonald, & Gerkey, 2005).

We compared our model data to two pools of human subject data: a meta-analysis of the simple ToM task (Wellman et al., 2001), and a more focused study of the more difficult ToM prediction task (Leslie et al., 2005). With respect to the simple ToM task, and considering age as the only variable, the model data had $r^2 = 0.51$ when compared to the linear regression model of (Wellman et al., 2001). This is considerably higher than their $r^2 = 0.39$ and approaches the R^2 of their multi-variate

model, 0.55. We are pleased with this result, especially since our model is a process model that learns to perform the task, and depends on fewer parameters.

For the second, more difficult ToM task, our model data matched the human data from (Leslie et al., 2005) which showed that only about 25% of children around 4.75 years of age have theory of mind of that level of sophistication. Although we were able to match the data, supporting our work, further experimental data are needed in order to truly distinguish our model from other possibilities.

	Intelligent		Natural	
	Mean	SD	Mean	SD
ToM	2.83	0.45	2.64	0.59
Simple Correction	2.03	0.51	2.01	0.69
Blindly Follow	1.14	0.35	1.34	0.59

(a) Rankings for the three conditions, where 3 is best.

	Intelligent		Natural	
	Mean	SD	Mean	SD
ToM	6.24	0.97	5.67	1.16
Simple Correction	5.00	1.15	4.64	1.27
Blindly Follow	2.81	1.60	3.28	1.70

(b) Ratings for the three conditions, where 7 is best.

Figure 4. Theory of mind results for both intelligence and naturalness.

3.5.3 Contribution to HRI Next, we used our ToM model to improve our robot's ability to interact with people. ToM is especially pertinent to this issue because research in psychology has shown that without ToM, people can be severely impaired in their abilities to interact naturally with others (Baron-Cohen, Leslie, & Frith, 1985). We adapted our ToM model for two robot scenarios: a patrol scenario and an office scenario. Each scenario presents two main situations where the human acts in an unexpected way, providing a total of four theory of mind opportunities. As an example of a ToM situation, in the patrol scenario, the robot and human start out with instructions to patrol the south area after they finish at their current location. Shortly after the scenario starts, however, they receive orders to change their next location to be the west area instead. After completing the current area, however, the human starts heading to the south. Using our approach, the robot infers that this is because the human forgot about the change in orders and corrects her. It did this by running simulations of different possible beliefs the human could have and realizing that the human could have remembered the wrong belief. It is then able to remind the human that the orders were updated, and they should head west. These scenarios were implemented on our MDS robot (Breazeal et al., 2008).

The key feature of this model, from an HRI perspective, is its fidelity to the human mind by its use of the entire ACT-R/E architecture during the ToM process. Other approaches have developed ToM architectures that include belief maintenance systems (Breazeal, Gray, & Berlin, 2009), infer human plans from their actions (Johansson & Suzic, 2005), or allow model ToM in terms of Markov random fields (Butterfield, Jenkins, Sobel, & Schwertfeger, 2009). These approaches, however, all place limiting assumptions on the human's behavior: that an agent can know exactly what another knows by observing them, or that inconsistencies in predicted versus observed behavior arise solely from differences in beliefs about the world. As we have argued above, however, humans are fallible and may act unpredictably because they forgot something, or because they may have matched the

wrong action to the current context because of the stochastic nature of their own neural machinery. Our ToM approach is able to overcome the limitations of the other approaches due to its use of cognitive models.

To test the advantages of theory of mind and our model in particular, we ran an experiment where we compared our ToM robot with two different ways of "understanding" humans: a robot that points out any discrepancy between what the human does and the robot's expected action (a simple correction robot); and a robot that simply follows the human around expecting the person to be correct (a blindly following robot). All three robots performed the two tasks above, with the simple correction robot and the blindly follow robot handling the ToM opportunities according to their respective conditions. Thirty-five participants viewed videos of a human confederate and the MDS robot interacting in these ways (see Figure 3) and were asked to evaluate the three robots' intelligence and naturalness comparatively (by ranking), and individually (by a Likert rating). Figure 4 show the results. Overall, participants ranked the ToM robot to be the best of the three; this difference was found to be significant (intelligence, Kruskal-Wallis $\chi^2(2, N = 35) = 86.1, p < 0.05$; naturalness $\chi^2(2, N = 35) = 59.8, p < 0.05$; Steel-Dwass post hoc pairwise statistic $ps < 0.05$ for all differences). Similarly, people rated the ToM robot to be both more natural and more intelligent than either the simple correction robot or the blindly following robot (intelligence, $\chi^2(2, N = 35) = 66.7, p < 0.05$, all $ps < .05$; naturalness, $\chi^2(2, N = 35) = 44.5, p < 0.05$, all $ps < .05$; Hiatt et al., 2011). A video, which includes a ToM interaction example, can be found at http://www.nrl.navy.mil/aic/iss/aas/movies/ep2-full-web-medium.mov (Hiatt, Harrison, Lawson, Martinson, & Trafton, 2011).

4. Discussion

In this paper, we described our work with ACT-R/E, an embodied computational cognitive architecture. ACT-R/E is a high-fidelity simulator for how people perceive, think, and act in the world. We described ACT-R/E's different modules, how each works separately, and how they are integrated together. We showed how the cognitive plausibility of ACT-R/E models can be evaluated by matching to different types of human data and we have introduced several models that have high plausibility. These models included finding the objects that people are looking at (Section 3.2), playing hide and seek with a person (Section 3.3), determining when someone has forgotten something in order to help them resume (Section 3.4), and understanding the beliefs of another person in order to help the robot and the person have a common understanding (Section 3.5).

As we have said earlier, a cognitive architecture is not the perfect tool for all robotic tasks. When optimal performance is desired (e.g., an answer that is fast and numerically accurate), ACT-R/E is assuredly not the ideal solution. When interacting with a person, however, our contention is that a cognitive architecture like ACT-R/E is one of the best tools available. There are several reasons that a cognitive architecture is a strong tool when interacting with people.

One contribution a cognitive architecture can make to HRI is a deep understanding of the processes and representations that people use (Trafton, Schultz, Cassimatis, et al., 2006). With knowledge of how people might perform in different situations, the robot can use that knowledge to improve overall performance. For example, suppose a robot investigates a burning building to find and save potential victims of the fire. Knowledge of where the victims are likely to go to protect themselves from fire and smoke given the cognitive, spatial and pragmatic constraints relevant to people can be very useful in finding potential victims.

Another reason for using a cognitive architecture (and ACT-R/E in particular) is to directly facilitate HRI by understanding people's specific knowledge, strengths and weaknesses. If a robot does not understand that people cannot see something if they are not looking at it, that people have fallible memories, or that people do not know everything that the robot (or other people) knows, then

interaction between the robot and the person will be decidedly awkward or may even have disastrous consequences. If, however, a robot knows when people are likely to forget something or is able to keep track of what a person knows and make inferences based on that knowledge, then the robot can use that information to create much more intelligent and natural interactions with the human.

In summary, our approach in this paper has been to show two main advantages of using an embodied cognitive architecture, and ACT-R/E in particular. First, it provides a deep understanding of people within the cognitive science tradition: our ACT-R/E models faithfully model people's behavior as they perceive, think about, and act on the world around them. Second, ACT-R/E models can use that deep understanding when dealing with the strengths, limitations, and knowledge of people. That understanding can be leveraged to leave the person alone when they are doing something they are good at, to help them when they need help and to provide them relevant knowledge that they do not have. The road to good, embodied cognitive models has been and continues to be long, but the power it provides HRI researchers makes it well worth the effort.

Acknowledgments

This work was supported by the Office of Naval Research, grant numbers N0001412WX30002 and N0001411WX20516 to GT. The views and conclusions contained in this document do not represent the official policies of the US Navy.

References

Altmann, E. M. (2000, August). Memory in chains: Modeling primacy and recency effects in memory for order. In L. Gleitmann & A. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society.*

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive Science*, *26*(1), 39-83, http://dx.doi.org/10.1207/s15516709cog2601_2.

Altmann, E. M., & Trafton, J. G. (2007). Timecourse of recovery from task interruption: Data and a model. *Psychonomic Bulletin & Review*, *14*(6), 1079-1084, http://dx.doi.org/10.3758/BF03193094.

Anderson, J. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261–295, http://dx.doi.org/10.1016/S0022-5371(83)90201-3.

Anderson, J. (2007). *How can the human mind occur in the physical universe?* Oxford University Press, http://dx.doi.org/10.1093/acprof:oso/9780195324259.001.0001.

Anderson, J., Albert, M. V., & Fincham, J. M. (2005, August). Tracing problem solving in real time: fMRI analysis of the subject-paced tower of Hanoi. *Journal of Cognitive Neuroscience*, *17*, 1261–1274, http://dx.doi.org/10.1162/0898929055002427.

Anderson, J., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036-60, http://dx.doi.org/10.1037/0033-295X.111.4.1036.

Anderson, J., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341–380, http://dx.doi.org/10.1006/jmla.1997.2553.

Anderson, J., Qin, Y., Sohn, M., Stenger, V., & Carter, C. (2003). An information-processing model of the bold response in symbol manipulation tasks. *Psychonomic Bulletin & Review*, *10*(2), 241–261.

Anderson, J., Reder, L., & Lebiere, C. (1996, August). Working Memory. *Cognitive Psychology*, *30*(3), 221–256.

Ashby, F., & Waldron, E. (2000). The neuropsychological bases of category learning. *Current Directions in Psychological Science*, *9*(1), 10–14, http://dx.doi.org/10.1111/1467-8721.00049.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*, 37–46, http://dx.doi.org/10.1016/0010-0277(85)90022-8.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(04), 577–660, http://dx.doi.org/10.1017/S0140525X99002149.

Barsalou, L. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science*, *2*(4), 716–724, http://dx.doi.org/10.1111/j.1756-8765.2010.01115.x.

Biederman, I. (1993). Geon theory as an account of shape recognition in mind and brain. *Irish Journal of Psychology*, *14*(3), 314-327, http://dx.doi.org/10.1080/03033910.1993.10557936.

Breazeal, C., Gray, J., & Berlin, M. (2009). An embodied cognition approach to mindreading skills for socially intelligent robots. *International Journal of Robotics Research*, *28*(5), 656-680, http://dx.doi.org/10.1177/0278364909102796.

Breazeal, C., Siegel, M., Berlin, M., Gray, J., Grupen, R., Deegan, P., et al. (2008). Mobile, dexterous, social robots for mobile manipulation and human-robot interaction. In *SIGGRAPH '08: ACM SIGGRAPH 2008 new tech demos.* New York, New York, http://dx.doi.org/10.1145/1401615.1401642.

Brooks, R., & Mataric, M. (1993). Real robots, real learning problems. In J. Connell & S. Mahadevan (Eds.), *Robot learning* (p. 193-213). Kluwer Academic Press, http://dx.doi.org/10.1007/978-1-4615-3184-5_8.

Butterfield, J., Jenkins, O. C., Sobel, D. M., & Schwertfeger, J. (2009). Modeling aspects of theory of mind with Markov random fields. *International Journal of Social Robotics*, *1*(1), 41-51, http://dx.doi.org/10.1007/s12369-008-0003-1.

Butterworth, G., & Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, *9*, 55-72, http://dx.doi.org/10.1111/j.2044-835X.1991.tb00862.x.

Byrne, M., & Anderson, J. (1998). Perception and action. In J. Anderson & C. Lebiere (Eds.), *The atomic components of thought* (pp. 167–200). Mahwah, NJ: Lawrence Erlbaum.

Cangelosi, A., & Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive Science*, *30*(4), 673–689, http://dx.doi.org/10.1207/s15516709cog0000_72.

Cassimatis, N. L. (2002). *A cognitive architecture for integrating multiple representation and inference schemes*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Cassimatis, N. L., Bello, P., & Langley, P. (2008). Ability, breadth and parsmiony in computational models of higher-order cognition. *Cognitive Science*, *32*(8), 1304-1322.

Cassimatis, N. L., Trafton, J. G., Bugajska, M. D., & Schultz, A. C. (2004). Integrating cognition, perception and action through mental simulation in robots. *Journal of Robotics and Autonomous Systems*, *49*(1-2), 12-23, http://dx.doi.org/10.1016/j.robot.2004.07.014.

Collett, T. H. J., MacDonald, B. A., & Gerkey, B. P. (2005). Player 2.0: Toward a practical robot programming framework. In *Proceedings of the Australasian Conference on Robotics and Automation.*

Corkum, V., & Moore, C. (1998). The origins of joint visual attention in infants. *Developmental Psychology*, *34*(1), 28-38, http://dx.doi.org/10.1037//0012-1649.34.1.28.

Craighero, L., Fadiga, L., Umiltà, C., & Rizzolatti, G. (1996). Evidence for visuomotor priming effect. *Neuroreport*, *8*(1), 347–349, http://dx.doi.org/10.1097/00001756-199612200-00068.

Diaper, D., & Waelend, P. (2000). World wide web working whilst ignoring graphics: Good news for web page designers. *Interacting with Computers*, *13*(2), 163–181, http://dx.doi.org/10.1016/S0953-5438(00)00036-9.

Doniec, M., Sun, G., & Scassellati, B. (2006). Active learning of joint attention. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robotics.* http://dx.doi.org/10.1109/ICHR.2006.321360.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *47*(6), 381–391, http://dx.doi.org/10.1037/h0055392.

Fong, T., Kunz, C., Hiatt, L. M., & Bugajska, M. (2006). The human-robot interaction operating system. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction.* http://dx.doi.org/10.1145/1121241.1121251.

Friedlander, D., & Franklin, S. (2008). LIDA and a theory of mind. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Proceedings of the 2008 Conference on Artificial General Intelligence* (p. 137-148). IOS Press.

Fu, W.-T., & Anderson, J. (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General*, *135*(2), 184–206, http://dx.doi.org/10.1037/0096-3445.135.2.184.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493-501, http://dx.doi.org/10.1016/S1364-6613(98)01262-5.

González, V., & Mark, G. (2004). Constant, constant, multi-tasking craziness: Managing multiple working

spheres. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 113–120). http://dx.doi.org/10.1145/985692.985707.

Gunzelmann, G., Anderson, J., & Douglass, S. (2004). Orientation tasks with multiple views of space: Strategies and performance. *Spatial Cognition and Computation*, *4*(3), 207–253.

Gunzelmann, G., Byrne, M. D., Gluck, K. A., & Moore, L. R. (2009, August). Using computational cognitive modeling to predict dual-task performance with sleep deprivation. *Human Factors*, *51*(2), 251–260, http://dx.doi.org/10.1177/0018720809334592.

Harrison, A. M., & Schunn, C. D. (2002). ACT-R/S: A computational and neurologically inspired model of spatial reasoning. In *Proceedings of the 24th Annual Meeting of the Cognitive Science Society.*

Harrison, A. M., & Trafton, J. G. (2010). Cognition for action: an architectural account for "grounded interaction". *Proceedings 32nd Annual Conference of the Cognitive Science Society*.

Hiatt, L. M., Harrison, A. M., Lawson, W. E., Martinson, E., & Trafton, J. G. (2011). Robot secrets revealed: Episode 002. In *Association for the advancement of artificial intelligence video competition.*

Hiatt, L. M., Harrison, A. M., & Trafton, J. G. (2011). Accommodating human variability in human-robot teams through theory of mind. In *Proceedings of the International Joint Conference on Artificial Intelligence.*

Hiatt, L. M., & Trafton, J. G. (2010). A cognitive model of theory of mind. In *Proceedings of the International Conference on Cognitive Modeling.*

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to automata theory, languages, and computation*. Reading, MA: Addison-Wesley.

Huttenlocher, J., & Kubicek, L. (1979). The coding and transformation of spatial information. *Cognitive Psychology*, *11*, 375-394, http://dx.doi.org/10.1016/0010-0285(79)90017-3.

Johansson, L. R. M., & Suzic, R. (2005). Particle filter-based information acquisition for robust plan recognition. In *Proceedings of the eighth international conference on information fusion.*

Jola, C., & Mast, F. (2005). Mental Object Rotation and Egocentric Body Transformation: Two Dissociable Processes? *Spatial Cognition & Computation*, *5*(2&3), 217–237.

Kawamura, K., Nilas, P., Muguruma, K., Adams, J. A., & Zhou, C. (2003). An agent-based architecture for an adaptive human-robot interface. In *Proceedings of the 36th Hawaii International Conference on System Sciences.* http://dx.doi.org/10.1109/HICSS.2003.1174288.

Kennedy, W. G., Bugajska, M. D., Harrison, A. M., & Trafton, J. G. (2009). "Like-me" simulation as an effective and cognitively plausible basis for social robotics. *International Journal of Social Robotics*, *1*(2), 181-194.

Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction*, *12*(4), 391-438.

Kortenkamp, D., Burridge, R., Bonasso, R. P., Schreckenghost, D., & Hudson, M. B. (1999). An intelligent software architecture for semiautonomous robot control. In *Autonomy Control Software Workshop, Autonomous Agents 99* (pp. 36–43).

Laird, J. E. (2012). *The soar cognitive architecture*. The MIT Press.

Lee, F., & Gamard, S. (2003). Hide and seek: Using computational cognitive models to develop and test autonomous cognitive agents for complex dynamic tasks. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (p. 1372). Boston, MA.

Lehman, J. F., Laird, J., & Rosenbloom, P. (2006). *A gentle introduction to soar: 2006 update* (Tech. Rep.). University of Michigan.

Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, *50*, 45-85, http://dx.doi.org/10.1016/j.cogpsych.2004.06.002.

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, *24*, 603-613, http://dx.doi.org/10.1348/026151005X55370.

Montemerlo, M., & Thrun, S. (2003). Simultaneous localization and mapping with unknown data association using fastSLAM. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).* http://dx.doi.org/10.1109/ROBOT.2003.1241885.

Mou, W., McNamara, Valiquette, C., & Rump, B. (2004). Allocentric and egocentric updating of spatial memories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 142–157, http://dx.doi.org/10.1037/0278-7393.30.1.142.

Mumm, J., & Mutlu, B. (2011). Human-robot proxemics: physical and psychological distancing in human-robot interaction and psychological distancing in human-robot interaction. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (pp. 331–338). http://dx.doi.org/10.1145/1957656.1957786.

Nagai, Y., Hosoda, K., Morita, A., & Asada, M. (2003). A constructive model for the development of joint attention. *Connection Science*, *15*(4), 211-229, http://dx.doi.org/10.1080/09540090310001655101.

Newcombe, N., & Huttenlocher, J. (1992). Children's early ability to solve perspective taking problems. *Developmental Psychology*, *28*, 654-664, http://dx.doi.org/10.1037//0012-1649.28.4.635.

Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.

Pezzulo, G., Barsalou, L., Cangelosi, A., Fischer, M., McRae, K., & Spivey, M. (2011). The mechanics of embodiment: A dialog on embodiment and computational modeling. *Frontiers in Psychology*, *2*(5), http://dx.doi.org/10.3389/fpsyg.2011.00005.

Ratwani, R. M., & Trafton, J. G. (2009). Developing a predictive model of postcompletion errors. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

Ratwani, R. M., & Trafton, J. G. (2011). A real-time eye tracking system for predicting and preventing postcompletion errors. *Human–Computer Interaction*, *26*(3), 205–245.

Reason, J. (1990). *Human error*. Cambridge University Press.

Reifers, A., Schenck, I. N., & Ritter, F. E. (2005, August). Modeling pre-attentive visual search in ACT-R. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society.* Mahwah, NJ.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358-367, http://dx.doi.org/10.1037//0033-295X.107.2.358.

Saint-Cyr, J., Taylor, A., & Lang, A. (1988). Procedural learning and neostyrial dysfunction in man. *Brain*, *111*(4), 941–960, http://dx.doi.org/10.1093/brain/111.4.941.

Salvucci, D., & Taatgen, N. A. (2008). Threaded Cognition: An Integrated Theory of Concurrent Multitasking. *Psychological Review*, *115*(1), 101–130, http://dx.doi.org/10.1037/0033-295X.115.1.101.

Satake, S., Kanda, T., Glas, D., Imai, M., Ishiguro, H., & Hagita, N. (2009). How to approach humans?: strategies for social robots to initiate interaction. In *Proceedings of the 4th International Conference on Human-Robot Interaction* (pp. 109–116). http://dx.doi.org/10.1145/1514095.1514117.

Schneider, D., & Anderson, J. (2011). A memory-based model of Hick's law. *Cognitive Psychology*, *62*(3), 193–222, http://dx.doi.org/10.1016/j.cogpsych.2010.11.001.

Sellner, B., Heger, F. W., Hiatt, L. M., Simmons, R., & Singh, S. (2006). Coordinated multi-agent teams and sliding autonomy for large-scale assembly. *Proceedings of the IEEE*, *94*(7), 1425-1444.

Spira, J., & Feintuch, J. (2005). The cost of not paying attention: How interruptions impact knowledge worker productivity. *Report from Basex*.

Szafir, D., & Mutlu, B. (2012). Pay attention! Designing adaptive agents that monitor and improve user engagement. In *Proceedings of the 31st ACM/SigCHI Conference on Human Factors in Computing* (pp. 11–20). http://dx.doi.org/10.1145/2207676.2207679.

Taatgen, N. A., & Dijkstra, M. (2003, August). Constraints on generalization: Why are past-tense irregularization errors so rare? In *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 1146–1151). Mahwah, NJ.

Taatgen, N. A., & Lee, F. J. (2003, August). Production Compilation. *Human Factors*, *45*(1), 61–76, http://dx.doi.org/10.1518/hfes.45.1.61.27224.

Taatgen, N. A., Rijn, H. van, & Anderson, J. (2007, August). An integrated theory of prospective time interval estimation. *Psychological Review*, *114*(3), 577–598, http://dx.doi.org/10.1037/0033-295X.114.3.577.

Tikhanoff, V., Cangelosi, A., & Metta, G. (2011). Integration of speech and action in humanoid robots: iCub simulation experiments. *IEEE Transactions on Autonomous Mental Development*, *3*(1), 17–29m http://dx.doi.org/10.1109/TAMD.2010.2100390.

Trafton, J. G., Altmann, E. M., Brock, D. P., & Mintz, F. E. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human Computer Studies*, *58*(5), 583-603, http://dx.doi.org/10.1016/S1071-5819(03)00023-5.

Trafton, J. G., Altmann, E. M., & Ratwani, R. M. (2011). A memory for goals model of sequence errors. *Cognitive Systems Research*, *12*, 134-143, http://dx.doi.org/10.1016/j.cogsys.2010.07.010.

Trafton, J. G., Cassimatis, N. L., Bugajska, M. D., Brock, D. P., Mintz, F. E., & Schultz, A. C. (2005). Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man, and Cybernetics*, *35*(4), 460-470, http://dx.doi.org/10.1109/TSMCA.2005.850592.

Trafton, J. G., & Harrison, A. M. (2011). Embodied spatial cognition. *Topics in Cognitive Science*, *3*(4), 686–706.

Trafton, J. G., Jacobs, A., & Harrison, A. M. (2012). Building and verifying a predictive model of interruption resumption. *Proceedings of the IEEE*, *100(3)*, 648–659, http://dx.doi.org/10.1109/JPROC.2011.2175149.

Trafton, J. G., Schultz, A. C., Cassimatis, N. L., Hiatt, L. M., Perzanowski, D., Brock, D. P., et al. (2006). Communicating and collaborating with robotic agents. In R. Sun (Ed.), *Cognition and multi-agent interaction: From cognitive modeling to social simulation* (p. 252-278). New York, NY: Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511610721.011.

Trafton, J. G., Schultz, A. C., Perzanowski, D., Adams, W., Bugajska, M. D., Cassimatis, N. L., et al. (2006). Children and robots learning to play hide and seek. In A. Schultz & M. Goodrich (Eds.), *Proceedings of the 2006 ACM conference on Human-Robot Interaction.* http://dx.doi.org/10.1145/1121241.1121283: ACM press.

Triesch, J., Teuscher, C., Deak, G., & Carlson, E. (2006). Gaze following: why (not) learn it? *Developmental Science*, *9*(2), 125-147, http://dx.doi.org/10.1111/j.1467-7687.2006.00470.x.

Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology-Human Perception and Performance*, *24*(3), 830-846, http://dx.doi.org/10.1037//0096-1523.24.3.830.

Tucker, M., & Ellis, R. (2001). The potentiation of grasp types during visual object categorization. *Visual Cognition*, *8*(6), 769-800.

Wellman, H. W., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655-684, http://dx.doi.org/10.1111/1467-8624.00304.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625-636, http://dx.doi.org/10.3758/BF03196322.

Authors' names and contact information:

Greg Trafton, Naval Research Laboratory, Washington, DC, USA
Email: greg.trafton@nrl.navy.mil

Laura Hiatt, Naval Research Laboratory, Washington, DC, USA
Email: laura.hiatt@nrl.navy.mil

Anthony Harrison, Naval Research Laboratory, Washington, DC, USA
Email: anthony.harrison@nrl.navy.mil

Frank Tamborello, Naval Research Laboratory, Washington, DC, USA
Email: frank.tamborello.ctr@nrl.navy.mil

Sangeet Khemlani, Naval Research Laboratory, Washington, DC, USA
Email: khemlani@aic.nrl.navy.mil

Alan Schultz, Naval Research Laboratory, Washington, DC, USA
Email: alan.schultz@nrl.navy.mil